Samananthar Dataset

- Samanantar dataset which is the largest publicly available parallel corpora collection for Indic languages.

- Below is the link provide to download the parallel corpus data of required English-X (indic-language) :

  https://indicnlp.ai4bharat.org/samanantar/

- In our work we considered the below language pairs :
  - English-Hindi (en-hi) : This contains 8.56 million sentence pairs.
  - English-Telugu (en-te) : This contains 4.82 million sentence pairs.

- Before getting into the wider spectrum of work we performed two levels of filtration over the 2 language pair corpora to get the filtered data.

  - Filtration levels :
    - Level 1 : Extraction of sentences between sentence (word) length of threshold 5 to 100. Below are the number of sentences extracted in each language pair.

      - English-Hindi (en-hi) : 79,61,610 sentences
      - English-Telugu (en-te) : 30,36,823 sentences

    - Level 2 : Removing the corresponding parallel sentences from both English-Hindi (en-hi) files which has english text in hindi file and junk in both english text file and hindi text file.

      - Retained english & hindi sentences : 76,22,024 sentences
      - Deleted english & hindi sentences : 3,39,586 sentences

    - Similarly Removing the corresponding parallel sentences from both English-Hindi (en-hi) files which has english text in hindi file and junk in both english text file and hindi text file.

      - Retained english & telugu sentences : 29,68,429 sentences
      - Deleted english & telugu sentences : 68,394 sentences